A literature review on multimodal deep learning models for detecting mental disorders in conversational data: Pre-transformer and transformer-based approaches

Zilei Shao

Harvey Mudd College, Claremont, 91711, USA

zoshao@g.hmc.edu

Abstract. This paper provides a comprehensive review of multimodal deep learning models that utilize conversational data to detect mental health disorders. In addition to discussing models based on the Transformer, such as BERT (Bidirectional Encoder Representations from Transformers), this paper addresses models that existed prior to the Transformer, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The paper covers the application of these models in the construction of multimodal deep learning systems to detect mental disorders. In addition, the difficulties encountered by multimodal deep learning systems are brought up. Furthermore, the paper proposes research directions for enhancing the performance and robustness of these models in mental health applications. By shedding light on the potential of multimodal deep learning in mental health care, this paper aims to foster further research and development in this critical domain.

Keywords: mental disorder detection, conversational data analysis, multimodal deep learning system.

1. Introduction

Mental disorders, such as anxiety and depression, continue to be the primary global cause of burden, afflicting millions of people of all ages [1]. Despite the availability of treatment options, the prevalence and burden of mental disorders have remained relatively constant over time, with little effect on reducing their impact [1]. Early detection and intervention are essential for mitigating the negative impact of these conditions on the quality of life of individuals and preventing more severe long-term effects [2]. Deep learning techniques from artificial intelligence were recently developed to assist mental health practitioners in making decisions based on historical patient data, such as healthcare records, psychological information, and activity on social networks [3].

An intriguing area of research is the development of multimodal deep learning models for diagnosing mental disorders using interactive data [4]. These models utilize data from multiple sensory modalities, such as speech, text, and facial expressions, to better comprehend and recognize emotional communication patterns associated with mental health issues [5]. Multimodal deep learning models can be divided into two broad categories: Transformer-based models, which take advantage of the advanced capabilities of the Transformer architecture, and pre-Transformer models, which rely on earlier methods for deep learning such as long short-term memory (LSTM), CNN, and RNN.

This literature review focuses on the differences between pre-Transformer and Transformer-based techniques and provides a comprehensive overview of recent advances in multimodal deep learning systems for detecting mental disorders using conversational data. The author discusses the advantages and disadvantages of various models, focusing on how they might be used in mental health assessment and diagnosis, in addition to future research directions and challenges within this swiftly evolving area.

2. Pre-Transformer deep learning models for mental disorder prediction using conversational data

Section 2 provides an overview of pre-Transformer models used to detect mental disorders from conversational data. Prior to the development of Transformer-based models, RNNs, CNNs, and LSTM networks were extensively used in a variety of applications. The efficacy of these models in identifying mental disorders will be investigated and examples showing how they are currently employed in the field of mental health will be provided to evaluate diverse modalities, such as text, speech, and facial expressions. By examining these pre-Transformer models, this paper aims to establish a solid foundation for comprehending the creation and development of deep learning approaches in the context of mental health research.

2.1. Overview of pre-Transformer models and their categories

2.1.1. CNNs. CNNs are a family of machine learning algorithms designed primarily for processing grid-like input, such as photographs with significant spatial correlations between features. CNNs consist of numerous layers, including convolutional, pooling, and fully linked layers. A crucial component of a CNN is the convolutional layer, which executes convolution procedures on the input data. Each convolution process uses a group of learnable filters, or kernels, to identify particular characteristics or patterns in the data. Pooling layers are then used to reduce the spatial dimensions of the feature maps generated by the convolutional layers, thereby reducing the network's computational cost and improving its generalization ability. Typically, fully connected layers are employed at the network's endpoint for classification purposes. Figure 1 illustrates the architecture of a typical CNN, proposed by Krizhevsky et al. [6].

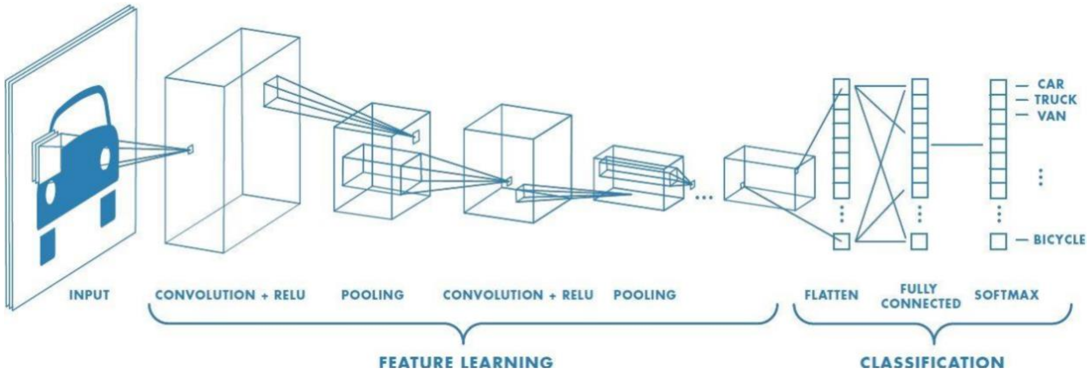

Figure 1. A typical CNN architecture. Initial layer neurons connect to a localized region of the input image, allowing them to capture local features [6].

As seen in Figure 1, the network consists of convolutional and pooling layers, with the convolutional layers learning spatial hierarchies and the pooling layers subsampling to reduce spatial dimensions. Towards the end of the architecture, multiple fully interconnected layers are used for advanced reasoning and feature integration. A softmax decision layer is then utilized to generate class probabilities, which facilitates classification tasks.

CNNs are able to learn hierarchical data representations that capture both low-level and high-level characteristics due to the combination of convolutional and pooling layers. In computer vision, CNNs are efficiently used for a variety of tasks such as semantic segmentation, object recognition, and picture categorization [6-8]. This is due to their ability to learn spatial hierarchies. CNNs can be used to assess visual data, such as facial expressions or body gestures, in the context of mental health detection to identify patterns and emotions associated with mental diseases [9].

2.1.2. RNNs and LSTM. RNN is a type of neural networks that have been developed to process sequential data. RNNs maintain internal states that allow them to comprehend temporal dependencies in the input data, as opposed to feedforward networks such as CNNs that analyze each input separately. This makes sequence-based tasks, such as time series analysis, natural language processing, and speech recognition, ideally suited for RNNs [10-12].

The architecture of an RNN consists of input, hidden, and output layers. At the hidden layer, recurrent connections that generate network loops allow data to persist over time steps. At each time step, the hidden layer receives both the current input data and the hidden state from the prior time step. On the basis of the current hidden state, the output layer makes a forecast. It has been demonstrated that the LSTM network, which is one of the most popular RNN types, is effective at learning long-term dependencies in data sequences. Memory cells, input gates, forget gates, and output gates are the components of the LSTM architecture. The gates control the passage of information through and out of memory cells, whereas memory cells retain data over an extended period. The input gate determines how much of the present input information is to be added to the memory cell, the forget gate determines how much of the memory cell's prior state is to be retained, and the output gate determines how much of the memory cell's content will be transmitted to the output. Figure 2 shows the structure of the memory block [13].

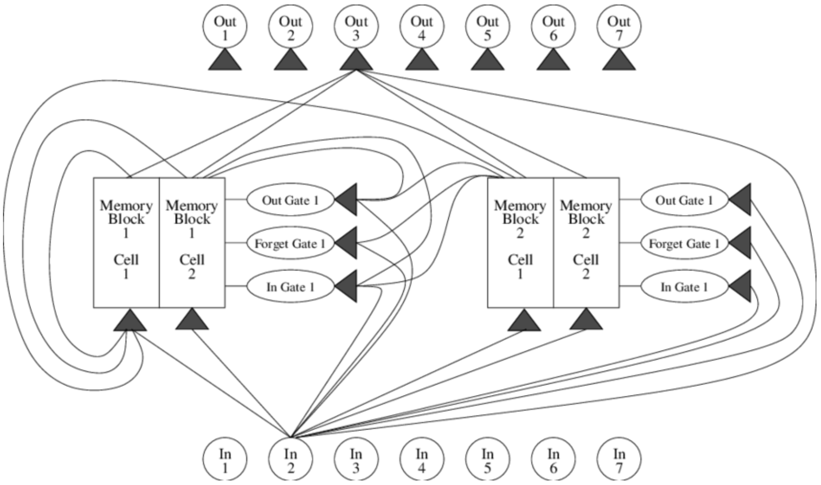

Figure 2. Three-layer LSTM architecture with recurrent connections limited to the hidden layer [13].

According to Figure 2, the hidden layer consists of four extended LSTM memory blocks, though only two are shown for clarity. Every memory block is composed of two memory cells. The diagram depicts a simplified view, revealing only a subset of the LSTM structure's connections. In the context of mental disorder detection, RNNs have been used to model the temporal dynamics of conversational data, such as speech signals and text transcripts, in order to capture and identify individuals' emotions [14].

2.2. Multimodal deep learning systems based on pre-transformer models

This section discusses multimodal deep learning systems that use conversational data to identify mental diseases using pre-Transformer models. Typically, these systems combine CNNs for processing visual data, such as body language and facial expressions, with RNNs or LSTMs for handling temporal data from speech signals and text transcripts. One system proposed by Malhotra et al. combines CNNs, RNNs, and LSTM networks using a multimodal deep learning-based architecture [15]. Several criteria, including accuracy, precision, recall, and F1-score, are used to evaluate the system's effectiveness. The results indicate that the proposed system is highly accurate at identifying depression and behaviors that may result in self-inflicted harm in social media posts. The structure of this system is shown in Figure 3 [15].

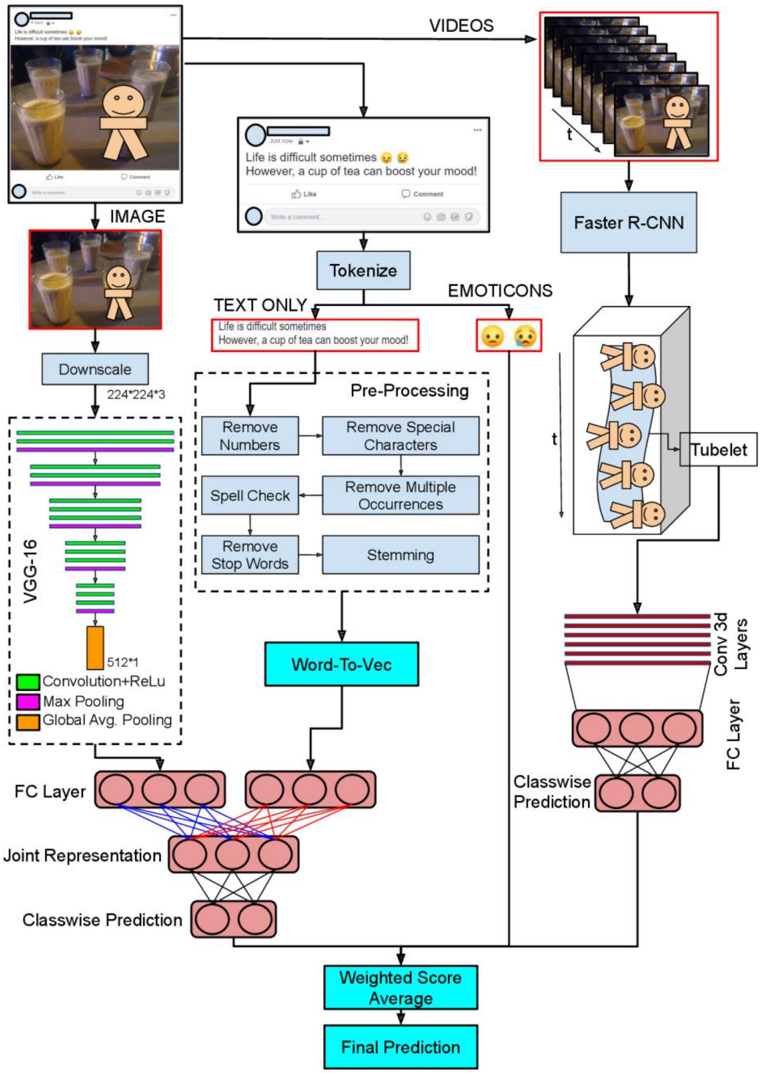

Figure 3. Architecture of the system proposed for multimodal examination of social media posts by Malhotra et al. [15].

Singhal et al. [16] proposed an alternative method for combining audio and text data using a Bidirectional Long Short-Term Memory (BiLSTM) model. The BiLSTM model is a variant of recurrent neural networks that processes input sequences both forward and backward. Audio and text data are preprocessed prior to being entered into distinct BiLSTM models. The outputs from each modality are then combined and routed via a thick layer for categorization purposes. BiLSTM model hyperparameters include the learning rate, the number of epochs, the optimization method, the loss function, and the activation function for the dense layer. The 74.3% accuracy and 5.14% error rate attained by the multimodal technique are superior to individual modalities, as demonstrated in the study of Rahul Singhal, et al. [16].

3. Transformer-based deep learning models for mental disorder prediction using conversational data

3.1. Overview of transformer models and their categories

This section will give a summary of Transformer models and their respective categories, with a particular emphasis on the original Transformer and its multi-head attention mechanism, as well as BERT.

3.1.1. Transformer model. Transformer models, a subset of deep learning architectures, have completely changed the natural language processing (NLP) industry and have now been applied to a number of other fields. These models use self-attention techniques to interpret input data, which enables them to effectively capture distant connections and intricate relationships. In several NLP tasks, transformer models have outperformed earlier state-of-the-art models, like RNNs and LSTMs.

The original Transformer, developed by Vaswani et al., consists of a multi-head self-attention mechanism that allows the model to concurrently attend to various parts of the input sequence [17]. The multi-head attention mechanism, which allows the model to simultaneously learn and focus on multiple elements of the input data, is an integral component of the Transformer architecture. Each attention head on the multi-head attention block is specialized in identifying specific relationships within the data as it simultaneously computes various attention mechanisms. As a result, the model is better able to comprehend intricate patterns and relationships between various points in the input sequence. Encoders and decoders are used for tasks such as machine translation and sequence-to-sequence learning in the original Transformer model. As shown in Figure 4, each encoder and decoder in the architecture consists of multiple identical layers. The encoder layers contain multi-head self-attention mechanisms and position-wise feed-forward networks, whereas the decoder layers contain a multi-head attention mechanism that attends to the encoder's output. Throughout the model, residual connections and layer normalization are utilized. Input and output embeddings incorporate positional encoding to preserve the order of the input sequence.

*3.1.2. BERT Model.*BERT is a significant variant of the Transformer model proposed by Devlin et al. [18]. It is an unsupervised, pre-trained language model that can be customized for a number of downstream applications, such as sentiment analysis, question answering, and named entity identification [18-20]. The primary innovation of BERT is its bidirectional training, which enables the model to learn context-aware word representations by analyzing input data in both directions. BERT's performance on a variety of NLP tasks is enhanced by its bidirectional training, which enables it to collect more precise contextual information.

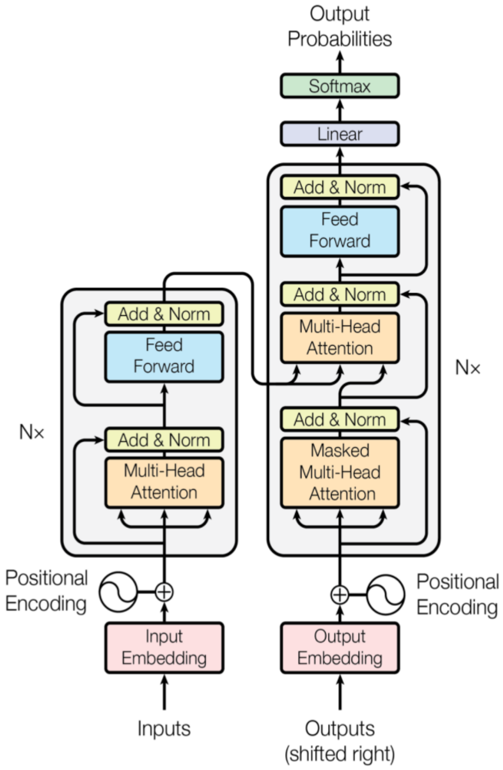

Figure 4. The architecture of the original Transformer model [17].

3.2. Multimodal deep learning systems incorporating transformer models

This section investigates the application of Transformer models to multimodal deep learning systems for the identification of mental health conditions, with a particular emphasis on BERT and multi-attention mechanisms. These systems combine different modalities, such as speech, text, and visual data, in order to improve performance and provide a more comprehensive understanding of the underlying mental states.

3.2.1. Utilizing BERT textual features. The depression detection model proposed by Makiuchi and colleagues uses a multiple-modal fusion of utterance and language representation [21]. The model employs a Gated Convolutional Neural Network (GCNN) followed by an LSTM layer for the speech modality and pulls deep broad spectrum information from a pretrained VGG-16 network. For textual embeddings, the model employs a CNN, which is followed by an LSTM layer. The input features for the fusion model are the concatenation of features acquired from the unimodal models.

The suggested multimodal model has the highest Concordance Correlation Coefficient (CCC) score of 0.403 for the DDS test partition. For the E-DAIC corpus development set, the CCC scores for the unimodal speech and language models were 0.497 and 0.608, respectively [21].

3.2.2. Utilizing multi-attention block. The Multi-attention Recurrent Network (MARN) is a complex deep neural network architecture with three primary components: the Long-short Term Hybrid Memory (LSTHM), the Multi-attention Block (MAB), and the output layer [22]. LSTHM is a common component that models the unique dynamics of each modality. It consists of a hybrid memory cell that combines data from both short-term and long-term memory cells, in addition to long-term memory cells. Each modality is assigned an LSTHM, allowing it to store crucial cross-view dynamics associated with it. The MAB component of MARN stores the temporal interactions between modalities in the hybrid memory of the LSTHM. Using attention mechanisms, the MAB focuses on different parts of the input sequence at each recurrent time step, thereby effectively capturing significant cross-modal interactions. The output layer of the MARN generates predictions for a variety of tasks, such as sentiment analysis, speaker attribute recognition, and emotion recognition [22]. The system's architecture is depicted in Figure 5 [22].

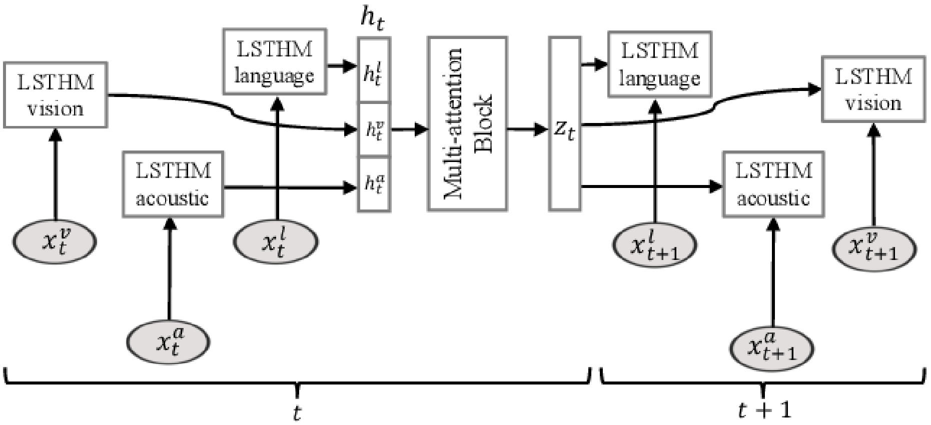

Figure 5. The architecture of the MARN, which consists of LSTHM and MAB components [22].

According to the paper, MARN has demonstrated leading-edge performance on six publicly available datasets for multimodal sentiment analysis, speaker trait recognition, and emotion recognition [22]. The CMU-MOSI dataset, which includes text, audio, and video modalities, yields the remarkable results.

4. Potential applications of multimodal deep learning models in mental health

Multimodal deep learning models have a great deal of application potential in a variety of mental health applications. Using data from multiple data sources, such as audio, text, and visual signals, these models can provide an in-depth depiction of a person's mental state. Some potential uses include:

4.1. Early diagnosis and detection

Multimodal deep learning models can be used to spot patterns and cues that are connected to mental diseases like depression or anxiety. These models could improve early diagnosis, resulting in prompt treatments and better outcomes by monitoring speech, facial expressions, body language, and text.

4.2. Telehealth and remote monitoring

As telehealth services are more widely used, multimodal deep learning models can be quite useful for remotely checking on patients' mental health. These models can assist clinicians in evaluating patients' mental states and monitoring their development over time by scrutinizing video conversations, texts, and other digital interactions. A system like this might allow clinicians to save time while also warning them when patients exhibit symptoms that the clinicians have missed.

4.3. Customized treatment recommendations

Multimodal deep learning models can help in the creation of individualized treatment programs by comprehending the special patterns and traits connected to a person's mental health. To meet the needs of the person, this may entail suggestions for counseling, modifications to a patient's medication, or self-care techniques.

Therefore, by providing prompt, precise, and individualized help, multimodal deep learning models have the potential to change mental health care. These models could have a big impact on how mental health concerns are identified, dealt with, and managed as they develop further.

5. Challenges facing in multimodal deep learning systems and possible next steps

A variety of multimodal deep learning methods for detecting mental health are covered in this study. Despite the encouraging outcomes, there are certain restrictions that call for more research:

5.1. Fusion techniques

The current multimodal models employ concatenation or late fusion, which are both rather straightforward fusion approaches. Exploring more complex fusion techniques, including attention-based or dynamic fusion mechanisms, may help better comprehend the connections between the different modalities and enhance the models' overall performance and robustness.

5.2. Visual feature learning

In the investigation in this paper, the visual models performed less well than other unimodal models in terms of results. Using cutting-edge methods for extracting visual features or using models that have already been trained specifically for facial emotion identification are two viable directions for improvement. This might result in a more realistic portrayal of visual signals, which are known to be crucial for assessing depression.

5.3. Models based on social media

The systems relying on social media data for mental health detection are constrained by users' willingness to disclose personal information and the possibility for errors when users do not openly communicate their feelings. The accuracy of these models could be increased by creating more sophisticated natural language processing methods to better comprehend the context and implicit emotions in user-generated content. Also, more users may volunteer their data for mental health assessment purposes if privacy-preserving mechanisms like federated learning or differential privacy are studied.

It may be possible to create multimodal deep learning systems for mental health detection and assistance that are more effective and robust by addressing these limitations and investigating these particular research directions.

6. Conclusion

This paper explores the potential of multimodal deep learning models for detecting and treating mental health disorders. In the context of mental health detection, the use of pre-Transformer models, such as CNNs and RNNs, and Transformer models, such as the original Transformer and BERT, as well as how they are combined to form multimodal deep learning systems are reviewed.

Despite the positive results, the field still faces a number of obstacles, including the need to investigate more advanced fusion techniques, enhance visual feature learning, and refine models based on social media data. By addressing these challenges and focusing on further research in these areas, it is possible to develop multimodal deep learning systems for mental health detection and assistance that are even more effective and robust.

Integration of advanced deep learning techniques into mental health care has the potential to revolutionize the identification, treatment, and management of mental health disorders. As these models continue to develop and mature, they can provide invaluable insights and tools for mental health professionals, patients, and society as a whole, resulting in improved outcomes and a greater understanding of mental health.

References

[1] GBD 2019 Mental Disorders Collaborators (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. The lancet. Psychiatry, 9(2), 137-150. https://doi.org/10.1016/S2215-0366(21)00395-3.

[2] Arango, C., Díaz-Caneja, C. M. and McGorry, P. D., et al. (2018). Preventive strategies for mental health. The lancet. Psychiatry, 5(7), 591-604. https://doi.org/10.1016/S2215-0366(18)30057-

9.

[3] Su, C., Xu, Z., Pathak, J. and Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. Translational psychiatry, 10(1), 116. https://doi.org/10.1038/s41398-020-0780-3.

[4] Tavabi, L. (2019). Multimodal Machine Learning for Interactive Mental Health Therapy. 2019 International Conference on Multimodal Interaction. https://doi.org/10.1145/3340555.3356095.

[5] Rabbi, M., Ali, S., Choudhury, T. and Berke, E. (2011). Passive and In-situ Assessment of Mental and Physical Well-being using Mobile Sensors. Proceedings of the ACM International Conference on Ubiquitous Computing. UbiComp (Conference), 2011, 385-394. https://doi.org/10.1145/2030112.2030164.

[6] Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60, 84-90. https://dl.acm.org/doi/10.1145/3065386.

[7] Girshick, R. B., Donahue, J., Darrell, T. and Malik, J. (2013). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 580-587.

[8] Shelhamer, E., Long, J. and Darrell, T. (2014). Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431-3440.

[9] Zhao, K., Chu, W. and Zhang, H. (2016). Deep Region and Multi-label Learning for Facial Action Unit Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3391-3399.

[10] Lipton, Z. C., Kale, D. C., Elkan, C. P. and Wetzel, R. C. (2015). Learning to Diagnose with LSTM Recurrent Neural Networks. CoRR, abs/1511.03677.

[11] Cho, K., Merrienboer, B. V. and Gülçehre, Ç., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing.

[12] Graves, A., Mohamed, A. and Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 6645-6649. https://doi.org/10.1109/icassp.2013.6638947.

[13] Gers, F. A., Schmidhuber, J. and Cummins, F. (1999). Learning to forget: continual prediction with LSTM. 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), Edinburgh, UK, 2, 850-855. doi: 10.1049/cp:19991218.

[14] Majumder, N., Poria, S. and Hazarika, D., et al. (2019). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6818-6825. https://doi.org/10.1609/aaai.v33i01.33016818.

[15] Malhotra, A. and Jindal, R. (1970). Multimodal Deep Learning Based Framework for Detecting Depression and Suicidal Behaviour by Affective Analysis of Social Media Posts: Semantic Scholar. EAI Endorsed Trans. Pervasive Health Technol. https://www.semanticscholar.org/paper/Multimodal-Deep-Learning-based-Framework-for-and-by-Malhotra-Jindal/fb81a94893616acc7697c047819cfc66ac712a09.

[16] Singhal, R., Srivatsan, S. and Panda, P. (2022). A Novel Multimodal Method for Depression Identification. https://doi.org/10.36548/jtcsst.2022.4.001.

[17] Vaswani, A., Shazeer, N.M. and Parmar, N., et al. (2017). Attention is All you Need. Advances in Neural Information Processing Systems (NIPS), 5998-6008.

[18] Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

[19] Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Lecture Notes in Computer Science, 194-206. https://doi.org/10.1007/978-3-030-32381-3_16.

[20] Akbik, A., Blythe, D. A. and Vollgraf, R. (2018). Contextual String Embeddings for Sequence

Labeling. International Conference on Computational Linguistics. ACL Anthology. https://aclanthology.org/C18-1139/.

[21] Makiuchi, M. R., Warnita, T., Uto, K. and Shinoda, K. (2019). Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. https://doi.org/10.1145/3347320.3357694.

[22] Zadeh, A., Liang, P. P. and Poria, S., et al. (2018). Multi-attention Recurrent Network for Human Communication Comprehension. Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, 32(1), 5642-5649. https://doi.org/10.1609/aaai.v32i1.12024.